**GOOGLE PLAY STORE APPLICATION CAPSTONE PROJECT REPORT**

**INTRODUCTION**

The data scraped from Google Play Store holds various exciting build and social information on android mobile applications. It can be used for both exploratory data analysis (EDA) and predicting how marketable an android mobile app is. The problem is to ascertain the impact of variables such as minimum required android version, price, size, ratings, as well as any other significantly impacting variables on the marketability (number of downloads) of an android mobile application. My prospective clients would be mobile application development companies. It is believed that an understanding of the significant factors that influence the marketability of a mobile application, would aid designers and developers in making informed decisions during application implementation and deployment. Based on data analysis, the effect of mobile application build and pricing on its marketability would be determined.

The important fields contained in the data set are Rating, Pricing, Size, Reviews, Category, AndroidVer (minimum required version), LastUpdate (date of last update), and Installs. It would have been nice to have the creation date for each observation, as this would have aided in determining if there was a significant relationship between creation date and number of installs.

**DATA WRANGLING EFFORT**

To clean the data, I had to look for data within each variable that needed to be changed or cleaned. To do this, I got a distinct set of each variable. A breakdown of my data wrangling efforts is listed below.

1. Confirmed distinct value of each variable before making any adjustments, this aided my wrangling efforts

2. Confirmed distinct values in the Category variable

- Changed "1.9"" in the Category variable to "NA"

3. Confirmed distinct values in the Ratings variable

- Converted "NaN" to "NA" in the Rating variable

4. Searched for any occurrence of alphanumeric values in the Reviews variable

- Changed each alphanumeric value to "NA"

5. Created the SizeKB variable from the Size variable:

- changed "varies with device" to "NA"

- changed "1000+" to "NA"

- removed all occurrences of "k" (kilobytes) from behind each size

- removed all occurrences of "M" (megabytes) from behind each size and multiplied the resultant value by 1000 (converted to kilobytes)

6. Removed all occurrences of "+" from behind each value in the Installs variable to make them numeric values

- Searched for all occurrences of letters in the Installs variable, and changed them to "NA"

- Created levels for Installs

7. Categorized number of installs into a new variable AdjustedInstall containing the values (extremely small, very small, small, medium, large, and very large)

- Created levels for AdjustedInstall

- Categorized number of installs into a new ordinal variable AdjustedInstall2 containing the values (1, 2 and 3)

- Created levels for AdjustedInstall2

8. Confirmed distinct values in the Type variable

- Changed all occurrences of "NaN" and "0" to "NA" in the Type variable

9. Confirmed distinct values in the Price variable

- removed all occurrences of "$" from the front of each Price value to make them numeric

- Changed all non-numeric values in the Price variable to "NA"

10. Changed all non-numeric values in the Prices variable to "NA"

11. Confirmed distinct values in the ContentRating variable

- Changed all empty values to "NA"

12. Confirmed distinct values in the Genres variable

- Changed "11-Feb-18" to "NA"

13. In the LastUpdated variable:

- got all values not containing the short month word format e.g. "Feb"

- got all values not containing the short month word format e.g. "Jan"

- changed "9.0.19" to "NA"

- split the LastUpdated variable into 3 columns and appended them to the data frame as "Day", "Month", and "Year"

- changed the empty value in the Month variable to "NA"

- changed the empty value in the Year variable to "NA"

14. Confirmed distinct values in the AndroidVer variable

- Changed the values "Varies with device", "NaN", and the blank value to "NA"

15. Categorized AndroidVer into a new variable MinimumVer containing the values (version 1s, version 2s, version 3s, version 4s, version 5s, version 6s, version 7s, version 8s)

- Created levels for MinimumVer

**EXPLORATORY DATA ANALYSIS**

The following are some of my initial observations after EDA;

* Most apps had downloads of 1,000,000 and above with an average number of downloads of 15,464,339
* There are 33 app categories; Family, Game and Tools as top 3 categories. The bottom 3 categories are Beauty, Parenting, and Comics. Among apps with extremely to very small downloads (1-10000), Family apps have the largest presence, followed by Tool apps, then Medical apps. Among apps with small to medium downloads (50000-500000), Family apps have the largest presence, followed by Gaming apps, then Tool apps. Among apps with large downloads (50000000-500000000), Gaming apps have the largest presence, followed by Family apps, then Communication apps. Among apps with very large downloads (1000000000+), Communication apps have the largest presence, followed by Social apps, then Gaming apps. However, overall, the communication apps have the highest average downloads, followed by Social apps, then Video\_Player apps. These 3 app categories have the highest average downloads
* The relationship between the number of app downloads and mean rating follows a curved non-linear model. The model shows that apps with small number of downloads from 5000 and below, had an increase in mean rating as the number of downloads reduced, this demonstrates that apps with downloads fewer than 5000 are negatively affected by ratings. On the other hand, the mean ratings of apps with downloads between 5,000 and 100,000,000 increase as the number of downloads increase, this illustrates that apps with this range of downloads are positively influenced by ratings. Also, the mean ratings of apps with downloads 100,000,000 and above decrease as the number of downloads increase, thus displaying that apps with 100,000,000+ downloads are negatively impacted by ratings.
* There is a strong positive relationship between average reviews and number of Installs. Thus, the mean reviews of apps increases as the number of downloads increases. Number of downloads is positively impacted by reviews
* The density of app size is right skewed with an average size of 21516.53. This shows that few apps had very large sizes. There is also a positive relationship between average app size and number of Installs. Thus, the mean size of apps increases as the number of downloads increases. App size seems to have a positive influence on number of downloads
* The density of app price is right skewed with an average size of 1.027368. This shows that most apps had zero to extremely low prices (in fact majority of the apps were free). There is also a slight negative relationship between average app price and number of Installs. Therefore, the higher the price, the lower the number of downloads. Price negatively affects number of downloads.
* Apps that can only be downloaded with android phones using a minimum OS version of 4 were the most downloaded, this was followed by apps requiring a minimum of OS version 2. It was also seen that apps that require a minimum of OS version 2 were downloaded most between 2011 and 2015, while downloads of apps requiring OS version 4 or higher took over in 2016, and grew exponentially all the way to 2018. This illustrates that phones with newer android versions can download more Google Play Store apps.
* The top performing app categories (99 percentile or more in Reviews, Rating and number of Installs) are Gaming apps, Communication apps, Family apps and Social apps.

**APPLYING MACHINE LEARNING TO DATA**

The research question at hand can be framed into a supervised classification machine learning problem. Supervised machine learning tasks, learns a function that maps an input to an output based on known and labelled data. However, because the outcome variable is ordinal, the ordinal logistic regression or regression tree method would be best suited for the problem. The regression tree method was finally picked, because the ordinal logistic regression model has issues with the Reviews variable. The research questions are as follows:

1. What is the probability that a small, medium or large number of downloads would be affected by consumer ratings, number of reviews, mobile application price, mobile application size, and mobile application minimum installation operating system version.

The outcome/ dependent variable (number of installs) was broken down into an ordinal variable (AdjustedInstall2) of 1, 2 and 3. Where 1 is small number of downloads (0 - 500), 2 is medium number of downloads (1,000 - 1,000,000), and 3 is large number of downloads (5,000,000 - 1,000,000,000).

The independent variables were minimum required mobile operating system (OS) for installation (MinimumVer), average consumer ratings (Rating), mobile app size in kilobytes (SizeKB), number of consumer reviews (Reviews), and mobile app price (Price). Regression Trees was used to find the impact strength of the independent variables on the independent variable. This was because the outcome/ dependent variable (AdjustedInstall2) was an ordinal variable of 3 values. The decision tree was split on the Reviews variable as the best fit for the model (number of reviews of 43000 and above would yield app downloads of 5,000,000 - 1,000,000,000 and above, number of reviews between 14 and 43000 would yield app downloads of 1,000 - 1,000,000, and number of reviews less than 14 would yield app downloads between 0 - 500). To evaluate the success of the technique used, the following steps were taken:

1. A training and testing set were created in the ratio 70:30 respectively. The training set was used to create a regression tree model.
2. The misclassification error rates in the training and testing set were compared, and they both were 11%.
3. Furthermore, a 10 fold cross validation was used in training the data, and the trained data was used in predicting the testing set for accuracy. The predicted accuracy was 91%.
4. Training controls of number =10 and repeats = 3 was used, as well as tuning grid parameters were used when cross validating the data.
5. The model had a p-value of 2.2e-16 which is very significant since p< 0.001

**FURTHER RESEARCH**

This project found number of reviews to be greatly significant in determining number of mobile app downloads. It would be nice to further research variables that influence number of customer reviews.

**RECOMMENDATIONS**

* The top performing app categories (99 percentile or more in Reviews, Rating and number of Installs) are Gaming apps, Communication apps, Family apps and Social apps. Therefore, mobile app developers should concentrate in developing applications within these categories.
* Free mobile apps had much more downloads than paid apps. I would recommend that mobile app developers deploy free apps that may have in-app purchases or/and adverts.
* EDA showed that apps with the most downloads were those that allowed mobile operating system backward compatibility. I recommend developers build apps that can be downloaded by users with at least android version 4.0.
* A review of 30 top applications on Google Play Store showed that they had 5,000,000 as minimum number of downloads, 4.1 as minimum rating and 40,000 as minimum number of reviews. This is in line with the EDA from this project, as well as the decision tree model prediction. Thus, to succeed, mobile app users should strive to build functional and user friendly apps within the gaming, communication, family and social categories.